WHAT BAYES HAS TO SAY ABOUT THE EVIDENCE PROCEDURE

David H. Wolpert and Charles E. M. Strauss

To appear in
"Proceedings of the 1993 Workshop on Maximum-entropy and Bayesian methods"
G. Heidbreder (Ed.). Kluwer Academics Publishers

Abstract

The "evidence" procedure for setting hyperparameters is essentially the same as the techniques of ML-II and generalized maximum likelihood. Unlike those older techniques however, the evidence procedure has been justified (and used) as an approximation to the hierarchical Bayesian calculation. We use several examples to explore the validity of this justification. Then we derive upper and (often large) lower bounds on the difference between the evidence procedure's answer and the hierarchical Bayesian answer, for many different quantities. We also touch on subjects like the close relationship between the evidence procedure and maximum likelihood, and the self-consistency of deriving priors by "first-principles" arguments that don't set the values of hyperparameters.

"... any inference must be based on strict adherence to the laws of probability theory, because any deviation automatically leads to inconsistency."
- S. Gull, in [5]

"(Some have) estimated alpha from the data and then proceeded as if alpha is known. It is better to use the standard methods of Bayesian statistics and integrate out alpha."
- B. D. Ripley, in [13]

1 Introduction

In many statistics problems one has one or more "hyperparameters" (sometimes called "nuisance parameters") which occur in the distributions of interest but may not be of direct interest themselves. Examples are a choice of model, a noise level, a regularization constant in a regression problem, and "α" in maxent image reconstruction.

How to deal with a hyperparameter? A full Bayesian approach is to marginalize out the hyperparameter. (This is "hierarchical Bayes" - see [1, 3].) A non-Bayesian approach might set the hyperparameter to a single value, and use that value throughout the subsequent analysis. For example, one might choose the hyperparameter via maximum likelihood - choose the hyperparameter γ such that the conditional probability $P(D \mid \gamma)$ (or alternatively $P(\gamma \mid D)$) is maximized, where D is one's data. Recently it has been claimed that this kind of non-Bayesian approach is a good approximation to the full Bayesian approach whenever $P(\gamma \mid D)$ is peaked as a function of γ [9, 11]. In the context of this claim, setting γ to the value maximizing $P(\gamma \mid D)$ is known as "the evidence procedure" [9, 11, 12, 14].

Even though the evidence procedure has become popular amongst some Bayesians, the validity of its claim to approximate the Bayesian approach has never been thoroughly discussed. Consequently the accuracy of the procedure as such an approximation is rarely checked or reported. Perhaps even more remarkably, for some applications the full Bayesian answer is easier to calculate and apply [16, 20, 3]. Yet many researchers jump straight to the approximation of the evidence procedure, without checking if the exact answer is tractable, or if not, if perhaps some approximation other than the evidence procedure is preferable.

In the first part of this paper we state the evidence procedure, giving both an intuitive argument that it is a good approximation and an intuitive argument that it is not. We then explore the validity of the procedure in a simple gaussians example. In this example the procedure fails miserably for certain objects of interest, but works for others. We end with a formal discussion giving lower and upper bounds on the approximation error incurred with the evidence procedure. The bounds concern error in evaluating the posterior at a point, in evaluating the full posterior (both supremum norm and L^n norm error), in estimating the predictive distribution, and in estimating expectation values. This discussion demonstrates explicitly that the naive justifications for the evidence procedure found in the literature are inadequate. It also has implications for the self-consistency of any "first-principles" argument for a prior that does not fix all hyperparameters in that prior.

A recurring theme throughout the paper is that for many quantities of interest, the evidence procedure becomes more accurate as the object of interest becomes more dominated by the likelihood distribution. In other words, for those quantities the procedure is most accurate when the prior is irrelevant, so that there is no need for Bayesian analysis.

We emphasize that here we only analyze how well the evidence procedure approximates the full Bayesian answer. We are not concerned with whether the procedure meets non-Bayesian desiderata. (E.g., desiderata like requiring that one's answer doesn't change when additional irrelevant information is introduced, or like the desiderata in section 6.5 of [11] that actually argue for the use of maximum likelihood in *all* contexts, not just those related to hyperparameters.) Nor do we make any claims concerning how one should use the posterior (e.g., take its mean vs. take its mode), an issue properly addressed by decision theory. Moreover, we make no claims about how well the procedure works in practice. (A procedure's being non-Bayesian does not mean it works poorly in practice.) Studies empirically comparing the evidence procedure to other methods for setting hyperparameters have given mixed results [7, 8, 13, 14, 16, 17, 18, 19, 20]. However in evidence's defense we note that MacKay has recently won a prediction competition [12] by using the evidence procedure, albeit in conjunction with some new techniques like stacking [2] and the use of different regularization hyperparameters for different parts of the space.

2 What is the evidence procedure?

To illustrate the evidence procedure, consider the case where the hyperparameter parameterizes the prior distribution over the hypothesis space of vectors f. (To distinguish it from the generic hyperparameter γ, this kind of hyperparameter is indicated by α.) Some examples are the MaxEnt and Gaussian distributions: $P(f \mid \alpha) \ = \ \exp(\alpha S(f))/Z_s(\alpha)$, and $P(f \mid \alpha) \ \propto \ \alpha^{N/2} \, e^{-\alpha |\vec{f}\,|^2}$, respectively.

Write the posterior distribution as

$$P(f \mid D) \ = \ \frac{1}{P(D)} \int P(\alpha, f, D) \; d\alpha. \tag{1}$$

Multiply and divide the integrand in (1) by $P(\alpha \mid D)$:

$$P(f \mid D) \ \propto \ \int \frac{P(\alpha, f, D)}{P(\alpha \mid D)} \; P(\alpha \mid D) \; d\alpha \ \propto \ \int P(f \mid \alpha, D) \; P(\alpha \mid D) \; d\alpha. \tag{2}$$

When $P(\alpha \mid D)$ is sharply peaked about α_{ev} it's natural to treat it as a delta function about α_{ev} and collapse the last integral in (2). The idea of collapsing Bayesian integrals this way is old, going back at least as far as [6]. It forms the conventional justification for the view that the evidence procedure is an approximation to the full Bayesian approach; the evidence procedure says that

$$P(f \mid D) \ \approx \ P(f \mid \alpha_{ev}, D) \ \propto \ P(f \mid \alpha_{ev}) \; P(D \mid f). \tag{3}$$

Under many circumstances (e.g., relatively flat $P(\alpha)$) this kind of reasoning also appears to support the idea of setting $P(f \mid D)$ to $P(f \mid D$, $\text{argmax}_\alpha P(D \mid \alpha))$, so long as $P(D \mid \alpha)$ is a peaked function of α. (In fact, this kind of reasoning appears to support setting α to the maximum of almost any distribution over α and D that is a peaked function of α.) So there is ambiguity in what peak we should set α to, i.e., in how to define α_{ev} (ambiguity that is reflected in the literature). Accordingly, when it's helpful for illustrative purposes, we will consider $P(D \mid \alpha)$ rather than $P(\alpha \mid D)$ and will take the term "evidence" to mean $P(D \mid \alpha)$ rather than (our default meaning) $P(\alpha \mid D)$.

Stripped of the context of equation (3), the idea of setting the hyperparameter to the value α_{ev} is essentially identical to the techniques of ML-II and generalized maximum likelihood [4, 1, 19]. The primary difference between the evidence procedure and those older techniques is that those older techniques do not attempt to justify themselves with the approximation in equation (3), but rather view setting $\alpha = \alpha_{ev}$ as *a priori* reasonable.

As it turns out, there are reasons to doubt the validity of equation (3). One such reason is that in general the change of variables $\alpha = \eta(\alpha')$ results in the evidence procedure returning $P(f \mid \alpha, D)$ for an α different from α_{ev}. That is, the Jacobian of the variable transformation can change the distribution's mode. (In other words, in general there will be functions η for which $P(\alpha' \mid D)$ is highly peaked about an α' which doesn't equal $\eta^{-1}(\alpha_{ev})$. For such an η the evidence procedure used with the hyperparameter α' returns a posterior distribution for f given by $P(f \mid \alpha'', D)$ where $\alpha'' \neq \alpha_{ev}$.) So the answer of the evidence procedure can change under a variable transformation of the hyperparameter, whereas the true posterior can not (cf. equation (1)). This suggests that the reasoning embodied in equations (1) through (3) must be flawed. More is needed than simply having a distribution over α and D that is a sharply peaked function of α.

Another reason to doubt the accuracy of the approximation in (3) arises from considering the evidence procedure from a graphical perspective. The contour plots in figure 1 show two hypothetical $P(\alpha, f \mid D)$'s, for one-dimensional f. The projections of these distributions onto the α and f axes are $P(\alpha \mid D)$ and $P(f \mid D)$, respectively. In both plots $P(\alpha \mid D)$ is peaked, about $\alpha = \alpha_{ev}$. The evidence procedure's posterior distribution is given by the slice of the original distribution through $\alpha = \alpha_{ev}$. In the left plot that slice resembles the true posterior projection. But in the right plot it does not. Again we see that $P(\alpha \mid D)$'s being peaked cannot be the sole criterion for the validity of the evidence approximation.

These problems are partially due to the fact that $P(\alpha \mid D)$ appeared in the integrand in (2) only after we multiplied and divided by it. So no matter how peaked the numerator $P(\alpha \mid D)$, it is exactly canceled by the denominator $P(\alpha \mid D)$. This suggests that the function $P(f \mid \alpha, D)$ appearing in equation (2) is just as rapidly varying a function of α as $P(\alpha \mid D)$, in which case collapsing the integral at α_{ev} is unjustified.

Note though that if the α-peak of $P(\alpha, f, D)$ is close to α_{ev}, there might be a fortuitous cancellation of peaks that renders $P(f \mid \alpha, D)$ a slowly varying function of α. (See equation (2).) While it is usually difficult to check whether precise cancellation occurs, at a minimum the peaks must overlap substantially for such cancellation to be possible. (This is proven formally in section five.) When there is such overlap it's possible that the evidence procedure closely approximates the Bayesian answer. Ironically, whereas the intuition behind equation (3) suggests that the procedure works better for more highly peaked $P(\alpha \mid D)$, the need for that narrow peak to overlap with the peak of $P(\alpha, f, D)$ suggests that the opposite is true.

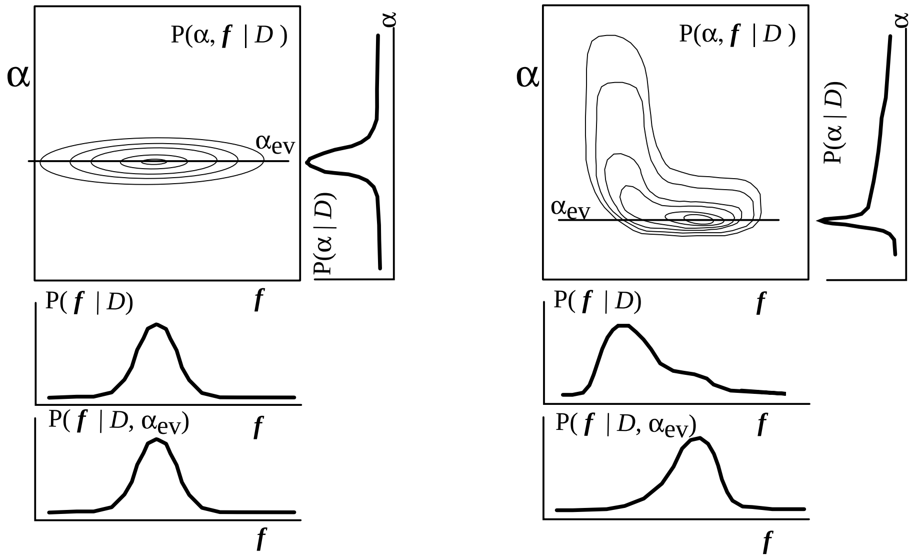

Figure 1: Contour sketches of hypothetical $P(\alpha, f \mid D)$'s along with their projections onto the α and f axes. The bottom plots are (proportional to) slices of the distributions through $\alpha = \alpha_{ev}$. The left sketch is a success of the evidence procedure, and the right a failure. The right sketch is similar to what one would get for the gaussian scenario discussed below.

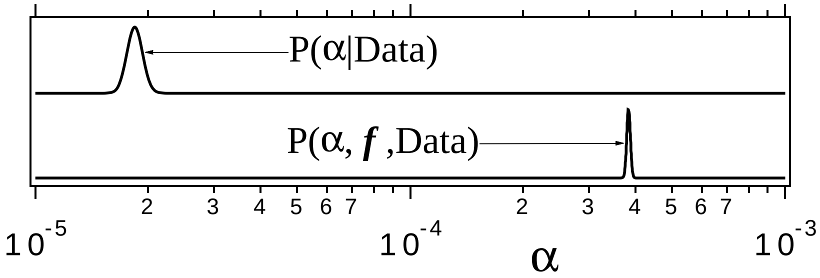

Figure 2: A comparison of $P(\alpha \mid D)$ and $P(\alpha, f, D)$ as functions of α shows they do not overlap. The data is taken from Gull's Susie reconstruction: f here is the MAP of the evidence procedure posterior f presented in Gull's article (see text).

(Theorem four below proves that that "opposite" is indeed true; the evidence procedure fails for almost all f in the regime of sufficiently peaked $P(\alpha \mid D)$.)

To illustrate this we consider Gull's famous Susie reconstruction [9]. Figure 2 plots $P(\alpha \mid D)$ and $P(f, \alpha, D)$ as functions of α for the f (i.e., the image) at the peak of the evidence procedure's posterior in Gull's Susie reconstruction. The two peaks clearly do not cancel, which means the argument leading to equation (3) does not hold. In addition, using Gull's Gaussian assumptions one can compute what f would have to be for the two peaks to overlap. This f corresponds to the peculiar images where $2\alpha S = N$; it is the image where the number of good degrees of freedom is the number of pixels.

It turns out that even when peaks cancel and $P(\alpha \mid D)$ is highly peaked, we still can't conclude that equation (3) is necessarily a good approximation. This is because $P(f \mid \alpha, D)$ need not be normalized over α, so the contribution to the integral from the (often very long) tails of the integrand in equation (2) can be as sizable as the contribution from around α_{ev}.

As a final example of the subtleties involved in equation (3) note that with enough hyperparameters the evidence procedure can produce a posterior that is highly peaked about the maximum likelihood f. (Nothing in the intuition behind equation (3) presumes α is low-dimensional. Indeed, some researchers have used the evidence procedure with high-dimensional α.) This follows from the equality $P(D \mid \vec{\alpha}) = \int df P(D \mid f) P(f \mid \vec{\alpha})$. This equality shows that for a sufficiently high-dimensional $\vec{\alpha}$ (i.e., sufficiently flexible $P(f \mid \alpha)$), to find the α maximizing $P(D \mid \vec{\alpha})$ one simply finds the α for which $P(f \mid \vec{\alpha})$ is highly peaked about the maximum likelihood f (i.e., about the mode of $P(D \mid f)$). Consequently, for that α, $P(f \mid D, \alpha)$ is also highly peaked about the maximum likelihood f.

3 The Gaussian distributions case

In this section we will focus on a particular example in which both the likelihood and the conditional prior distribution are gaussians. For simplicity the likelihood does not involve convolutions. The prior is centered on the origin and the likelihood is centered at a point D all of whose components have equal magnitude d. (These restrictions entail no loss of generality due to the translational and rotational invariance of gaussians). Accordingly, with N the dimension of f, the likelihood and (conditional) prior are given by

$$P(D \mid f) \;\propto\; \beta^{N/2}\, e^{-\beta\left|\vec{f}\,-\,D\right|^2}, \text{ and } P(f \mid \alpha) \;\propto\; \alpha^{N/2}\, e^{-\alpha\left|\vec{f}\;\right|^2} \qquad [4]$$

To agree with common usage, we will take the prior over α to equal $1/\alpha$ from α_{min} to α_{max} and zero elsewhere. We will be interested in the common case where α_{min} is very close to zero. Since our analysis won't depend on the exact value of α_{min} (the primary effect of that value is to set the overall normalization), here we will set it equal to 0. Also, for this section, we will treat α_{ev} as though it equaled $\text{argmax}_\alpha P(D \mid \alpha)$. It is straightforward to redo the analysis under different restrictions.

Evaluating $\int d\alpha P(f \mid \alpha)$ gives $P(f)$ in terms of the incomplete gamma function:

$$\begin{aligned} P(f) \;&\propto\; \tfrac{1}{|f\;|^N}\;\Gamma\left((N/2),\;\alpha_{\max}|f\;|^2\right) \\ &\approx\; \tfrac{1}{|f\;|^N} \quad \text{when } \alpha_{\max}\,|f\;|^2 \gg\; N/2. \end{aligned} \qquad [5]$$

Note that for f away from the origin, the prior falls off as a reciprocal power of distance from the origin; even though $P(f \mid \alpha)$ is gaussian $P(f)$ is not. (See theorem one below for a proof of the generality of this phenomenon.) Since the true posterior is proportional to the product of the prior with the likelihood, it too is non-gaussian. However the evidence procedure's posterior is gaussian, so the two posteriors must differ. To calculate the difference we must find the evidence procedure's posterior, and to do that we must first evaluate

$$P(D \mid \alpha) \;=\; \int df P(f, \alpha \mid D) \;\propto\; \left[\sqrt{\frac{\alpha\beta}{\alpha+\beta}}e^{-\frac{\alpha\beta}{\alpha+\beta}d^2}\right]^N. \qquad [6]$$

We can solve for the peak of this distribution, α_{ev}:

$$\alpha_{ev} \;=\; \frac{\beta}{2\beta d^2 - 1}. \qquad [7]$$

So the evidence procedure's posterior is a gaussian centered between the peaks of the prior and likelihood (i.e., between $f = 0$ and $f = d$):

$$P(f \mid D, \alpha_{ev}) \;\propto\; (\alpha_{ev}\beta)^{N/2}\, e^{-\beta|f\,-\,D|^2 - \alpha_{ev}|f\;|^2} \;\propto\; e^{-(\beta+\alpha_{ev})\left|f\,-\,\frac{\beta}{\alpha_{ev}+\beta}D\right|^2}. \qquad [8]$$

Note that d is the distance along any coordinate separating the peaks of the prior and the likelihood. Therefore $2\beta d^2$ is the separation between the peaks measured in units of the likelihood's width. But equation (7) only has a meaningful solution if $2\beta d^2 > 1$; unless the peaks are separated by more than the width of likelihood, there isn't a peak in the evidence. In this sense the evidence procedure is not even well-defined unless the data are unexpected. (We use the term "unexpected" a bit loosely here; more formally - and laboriously - one could analyze how "unexpected" the data are by considering the width of the prior predictive distribution rather than the width of the likelihood.) Moreover, as the separation increases beyond two widths, so that $2\beta d^2 > 2$, the value α_{ev} becomes smaller than β. Yet as α_{ev} shrinks below β the evidence procedure's approximation to the posterior approaches the likelihood distribution. So as we pass the condition allowing the evidence procedure to be well-defined, the data become more unexpected, and the evidence procedure produces a posterior which increasingly approximates the likelihood.

These and related effects are illustrated in figure 3. Since the evidence approximated posterior is a symmetric gaussian it is fully characterized by any single one-dimensional slice through its peak. This is not the case with the true posterior unfortunately, since that posterior is not symmetric about its peak. Nonetheless, we can learn a lot about the true posterior by looking at a slice through it going from the origin out along the D direction in f space. Figure 3 shows this slice and the corresponding slice of the evidence procedure's posterior for various separations, i.e., various values of $2\beta d^2$. The likelihood is also shown. The plots for other slice directions exhibit similar behavior.

These plots show that the evidence and true posteriors have different symmetries, peak positions and widths. Moreover the true posterior can have two peaks whereas the evidence procedure's posterior only has one, and the true posterior tends to have (sometimes much) more of its probability "mass" near the origin. Also note that the neither the peak position nor peak widths of the two distributions approach one another until the distributions start to converge on the likelihood - at which point the true posterior is about as well approximated by the likelihood as it is by the evidence procedure's posterior.

For large enough α_{max} and α_{min} close to 0, as N increases the peaks of the true posterior and of the evidence procedure's posterior don't move, nor does the position of the peak of the evidence move. But all those distributions—and in particular the plots in figure 3—become sharper (cf. equations (4, 5, and 8), and compare figures 3b and 3d). (Due to this sharpening of peaks the plots for high N values aren't very informative; this is why the plots are for low N values even though the evidence isn't very peaked for low N values.) So as N increase, the evidence becomes more peaked. But at the same time the discrepancy between the true posterior and the evidence procedure's posterior gets *worse*, not better.

Given all this, it seems fair to say that the evidence procedure's posterior is a poor representation of the true posterior—except for in the case when the prior doesn't matter (i.e., when things are likelihood dominated). Nonetheless, in some circumstances, the evidence

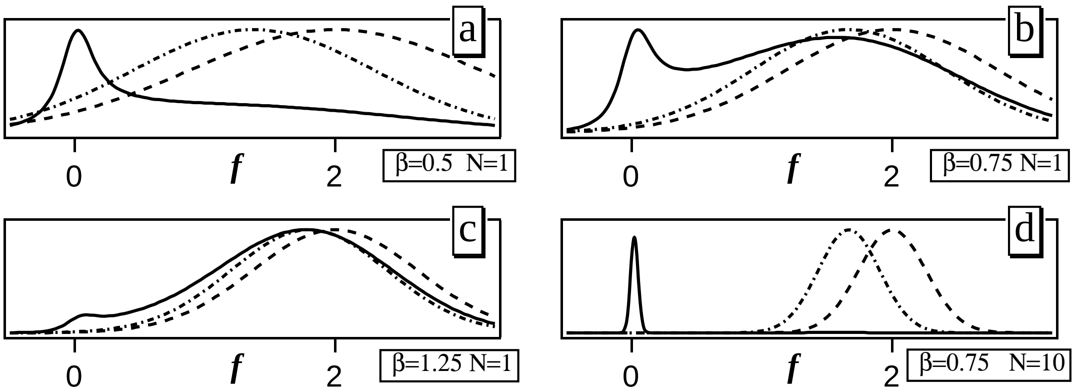

Figure 3: **Solid line**: True posterior, $P(f \mid D)$; **Dot-Dash**: Evidence procedure's posterior $P(f \mid \alpha_{ev}, D)$; **Dashed**: Likelihood $P(D \mid f)$. Going from figure (a) through (c), there is increasing distance (i.e., increasing $2\beta D^2$) between the peaks of the prior and the likelihood. For $2\beta d^2 < 1, \alpha_{ev}$ is undefined. Figure (d) increases the dimension from N = 1 to N = 10; the mismatch between the distributions becomes worse. ($\alpha_{max} = 100, d = 2$)

procedure's posterior could provide a good approximation for calculating low-dimensional expectation values. This will occur if erroneous behavior in the tails of the distribution "compensates" for erroneous behavior in the central regions. (See section 4 below.)

Finally, we point out that it is a simple matter to calculate the true prior (and therefore the posterior) not only when the conditioned prior is gaussian, but also when it is entropic (see equation (5) and [16]). Moreover, for both scenarios one can often directly approximate the exact posterior with a convenient form. Equation (5) presents an example of this for the gaussian prior case, and for the entropic prior such a direct approximation is $P(f) \sim 1/S(f)^{N/2}$, where S is the entropy (see [16]). Nonetheless, one can not rule out the possibility that there might be cases where the evidence procedure's functional form for the posterior is more convenient than "direct approximations" for the posterior. On the other hand of course, unlike the exact calculation's form for the posterior, generating the evidence procedure's form entails recalculating α_{ev} for each new data set.

4 Using evidence for things other than the posterior

Interestingly enough, all this doesn't mean that the evidence procedure is useless. This is because even though it gets the posterior wrong, *when certain conditions are met* the evidence procedure's approximation for low-dimensional expectation values can be excellent.

As an example, consider the posterior expected value of a function $g(f)$: $\langle g \rangle \equiv \int d\alpha \int df \; g(f) \; P(f, \alpha \mid D)$. Suppose that g is a simple function of a single coordinate f_j, and that $P(f, D \mid \alpha)$ factors as $\Pi_{k=1}^{N} P(f_k, D_k \mid \alpha)$ (as it does in our gaussians example). Then by equation (2),

$$\langle g \rangle = \int_{\alpha_{\min}}^{\alpha_{\max}} d\alpha \int df g(f_j) \frac{P(f, \alpha \mid D)}{P(\alpha \mid D)} P(\alpha \mid D).$$

Cancelling terms between the numerator $P(f, \alpha \mid D)$ and the denominator $P(\alpha \mid D) = \int df P(f, \alpha \mid D)$ (recall the assumption that $P(f, D \mid \alpha)$ factors), we see that

$$\langle g \rangle = \int_{\alpha_{\min}}^{\alpha_{\max}} d\alpha P(\alpha \mid D) R(\alpha) \qquad [9]$$

where $R(\alpha) \equiv \frac{\int df_j g(f_j) P(f_j, D_j \mid \alpha)}{\int df_j P(f_j, D_j \mid \alpha)} = \int df_j g(f_j) P(f_j \mid D_j, \alpha)$.

Equations (9) and (2) have the same form, except that in equation (9) the ratio occurring in the integrand ($R(\alpha)$) only involves one-dimensional quantities. As a result, often equation (9) does not give us the same difficulty that equation (2) did; since in equation (9) the denominator of the ratio is a one-dimensional integral, it is often not strongly peaked, so to have the ratio be smooth on the scale of the peak of the evidence does not require that the numerator of that ratio be strongly peaked, as it did in equation (2). So as long as: α_{max} is not too large (so that the tails don't contribute much); $R(\alpha)$ is not a rapidly varying function (a condition often met for simple expectation values like the mean); and $P(\alpha \mid D)$ is a highly peaked function of α (cf. equation (6)); then calculating the expected g by collapsing the integral over α down to the peak of $P(\alpha \mid D)$ might be justified.

$R(\alpha)$ and $P(\alpha \mid D)$ for the gaussians case are sketched in figure 4 for $g(f) = f$ (so $\langle g \rangle$ is the posterior average f). To highlight the important aspects of the plot, $P(\alpha)$ is flat between 0 and α_{max} rather than Jeffreys. These plots shows that slowly-varying $R(\alpha)$ and peaked $P(\alpha \mid D)$ is not uncommon, provided one has appropriate choices of α_{max} and the like. (Note that this is not the behavior of all the plots however. Also note the logarithmic scale of the x axis that "compresses" the tails.) So in some circumstances the evidence procedure can accurately estimate low-dimensional expectation values even if it poorly approximates the (high-dimensional) posterior distribution. To help understand this in light of the preceding discussion, note that $P(\alpha \mid D)$ is usually only highly peaked on the likelihood-dominated side of the midpoint in $R(\alpha)$. And of course in the likelihood-dominated regime we are free to introduce approximation error into the prior.

Note that all of this depends on the tails in figure 4 being relatively unimportant, which usually holds only if α_{max} is not too large. For example, in the gaussians case, for large enough α_{max} the tails of $P(\alpha \mid D)$ will provide more weight in the integral over α than the

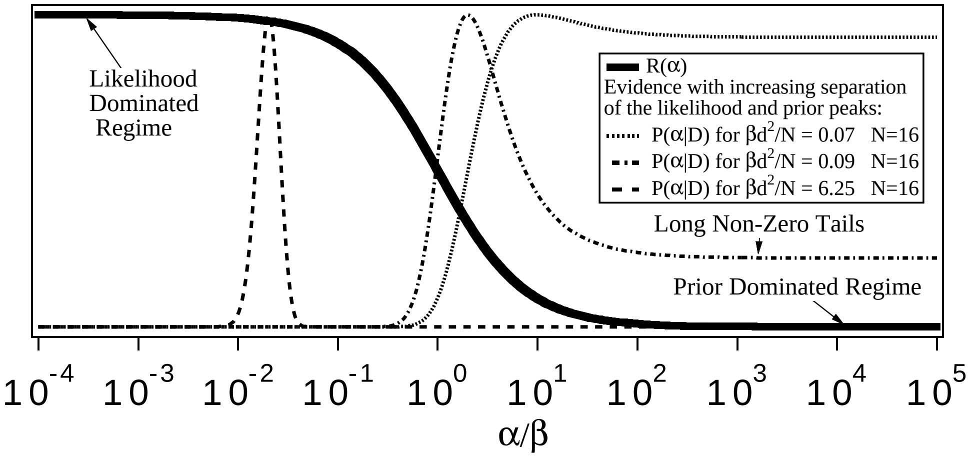

Figure 4: $R(\alpha)$ makes a smooth transition from the prior-dominated to the likelihood-dominated regime. It is weighted by $P(\alpha \mid D)$ in the integral giving $\langle g \rangle$. The long tails of $P(\alpha \mid D)$ can outweigh the peak of $P(\alpha \mid D)$ in the integral, particularly when that peak lies beyond the crossover point from the likelihood-dominated regime.

peak does. In such a situation, we are not justified in "collapsing the integral down to the peak", and the evidence's procedure's approximation for the expectation value is poor.

Unfortunately though, there is a lot of confusion about how to choose α_{max}. In particular, while a large α_{max} does indeed result in a less informative $P(\alpha)$, it results in a *more* informative $P(f)$. This is because the larger α_{max} is, the narrower $P(f)$ becomes. (Similar "conjugate" behavior in a different context has been discussed by Jaynes [10].) This is a special example of the following more general rule: if one knows the physical meaning of a hyperparameter, then one can set the prior over it directly, without concern for how that prior affects $P(f)$. However if the hyperparameter has no physical meaning, and if one sets the prior over it without taking into account how that prior affects $P(f)$, then one is introducing (usually fictitious) prior "knowledge" concerning the ultimate object of interest, f. This problem is particularly pronounced if $P(f \mid \alpha)$ is somewhat ad hoc, like in the case of neural nets, where f is an input-output mapping, and $P(\alpha)$ only sets $P(f)$ indirectly, by means of an intermediate distribution over "weight vectors" [21].

There are many other quantities of interest in addition to the posterior and its low-dimensional marginalizations. Two such quantities are the posterior over a single coordinate (i.e., $P(f_i \mid D)$) and the predictive distribution for new data given old data (i.e., $P(\text{new data set} = D' \mid D)$). Since the posterior over a single coordinate is a low-dimensional marginalization of the full posterior, we expect the evidence procedure to estimate it accurately when it estimates other low-dimensional marginalizations well. On the other hand, the predictive distribution is a high-dimensional object, and therefore we expect the evidence procedure to estimate it as poorly as it does the full posterior.

Yet another quantity of interest is the mode of the posterior, the "MAP" f. Since the MAP f is not a low-dimensional marginalization of the posterior, one would not expect the evidence procedure to approximate it well unless things are likelihood dominated. This is the case with gaussians for example - see figure 3.

Despite this though, applications of the evidence procedure frequently concentrate on the f-mode of $P(f \mid \alpha_{ev}, D)$. This isn't as unreasonable as it might seem if $P(f \mid \alpha_{ev}, D)$ is symmetric and unimodal, since for such a distribution the mode equals the mean. So when the evidence procedure's posterior is symmetric and unimodal, finding the mode of that posterior provides an accurate estimate of the true posterior's mean (if it so happens that the mean of the evidence procedure's posterior is a good approximation of the true posterior's mean - cf. equation (9)). We speculate that this is the origin of the cryptic claim that the evidence procedure estimates "where most of the mass is" correctly.

So in these symmetric and unimodal circumstances it is indeed sensible to concentrate on the mode of the evidence procedure's posterior. However when the evidence procedure's posterior is either asymmetric or multimodal, the peak of the procedure's posterior does not equal its mean. For such cases the mode of the procedure's posterior has no special significance, and there is no reason to concentrate on that mode. In particular, this problem affects use of the evidence procedure with the entropic prior, and with (highly multi-modal) neural nets. Ironically, these are two situations in which it happens to be particularly common for researchers to concentrate on modes of the evidence procedure's posterior.

As a final example of a quantity of interest, note that in many applications one is more concerned with unusual events than with likely events. (For example, a battleship's captain might not be interested in a "typical" reconstruction of a radar-image, but rather in the probability that that image was created by an approaching periscope.) In such a case we are interested in the probability distribution across the tails of the hypothesis space. However in general there is no reason to believe that the evidence procedure approximates such tails well. In particular, in the gaussians example the ratio of the true posterior to the evidence procedure's posterior goes to infinity in the tails of f (cf. equations (5, 8)). In the final analysis, whether or not a particular use of the evidence procedure is sound depends on what one wants to know (which in turn is determined by one's loss function).

5 Formal bounds on evidence's error

This section presents a formal analysis of upper and lower bounds on the error incurred by using the evidence procedure. (Some of these results correct deficiencies in the results reported in [20].) In most of this analysis we will not restrict attention to hyperparameters which occur in the conditional prior, so we denote hyperparameters by γ rather than α. Also, although most of this analysis goes through essentially unchanged when γ is multi-dimensional, for simplicity only the one-dimensional γ case is presented here.

This section is organized as follows. First it is proven that $P(f)$ can not be of the form $P(f \mid \gamma = \kappa)$ for some constant κ (i.e., marginalizing out a hyperparameter can never be equivalent to setting it to some particular value). It is argued that this means that "first-principles" arguments for a prior which don't set the value of the hyperparameter are not self-consistent. It also means that the evidence procedure will *always* have some error.

Next the reasoning of section 2 is formalized to derive an upper bound on the error of the evidence procedure. Like many of the other results presented in this section, this upper bound applies to a wide variety of possible uses of the evidence procedure.

Then it is shown that the separation between the γ-peaks of $P(f, \gamma \mid D)$ and $P(\gamma \mid D)$ must be small or the evidence procedure's error will be large (cf. the discussion of "fortuitous cancellation of peaks" near the end of section 2). This is done by both showing that the upper bound on the error increases with that separation, and then by deriving a lower bound on the error which increases with that separation. So by measuring the separation one can test the evidence procedure. In addition, the lower bound can be used to show that when $P(\gamma \mid D)$ is highly peaked—exactly the situation which traditionally was thought to justify the evidence procedure—the evidence procedure can give an accurate estimate of the entire posterior $P(f \mid D)$ only if that posterior is likelihood-dominated.

Finally, we discuss how well the evidence procedure performs when one uses error measures like the L^n difference between the correct posterior and the evidence procedure's guess for that posterior.

We start with a proof that for a broad class of $P(f \mid \gamma)$'s, there is no non-pathological scenario for which the evidence procedure's approximation to $P(f)$ is correct:

Theorem 1: Assume that for those γ for which it does not equal zero, $P(f \mid \gamma) \propto e^{-\gamma U(f)}$ for some function $U(.)$. Then the only way that one can have $P(f) \propto e^{-\kappa U(f)}$ for some constant κ is if $P(\gamma) = 0$ for all $\gamma \neq \kappa$.

Proof: Our proposed equality is $e^{-\kappa U} = \int d\gamma T(\gamma) \times e^{-\gamma U}$, where the integration limits are implicitly restricted to the region where $P(f \mid \gamma) \neq 0$, and where $T(\gamma) \equiv P(\gamma) \times \int df e^{-\kappa U(f)} / \int df e^{-\gamma U(f)}$. (Note that for both $P(f)$ and $P(f \mid \gamma)$ to be properly defined, both integrals in the definition of $T(.)$ must be greater than zero and finite.) We must find an κ and $T(\gamma)$ such that this equality holds for all realizable values of U. Let u be such a realizable value of U. Take the derivative with respect to U of both sides of the proposed equality t times, and evaluate for $U = u$. The result is $\kappa^t = \int d\gamma (\gamma)^t \times R(\gamma)$ for any integer

$t \geq 0$, where $R(\gamma) \equiv T(\gamma) \times e^{u(\kappa-\gamma)}$. Therefore $\int d\gamma (\gamma - \kappa)^2 \times R(\gamma)) = 0$. Since both $R(\gamma)$ and $(\gamma - \kappa)^2$ are nowhere negative, this means that for all γ for which $(\gamma - \kappa)^2 \neq 0$, $R(\gamma)$ must equal zero. Therefore $P(\gamma)$ must equal zero for all $\gamma \neq \kappa$. QED.

Theorem one has two important consequences. First, consider any "first principles" argument which says that the prior over f is proportional to $K(f)e^{-\gamma U(f)}$ for some $U(.)$ and $K(.)$ but does not fix γ. Our ignorance concerning γ implies a non-delta function distribution $P(\gamma)$. By theorem one, such a distribution ensures that $P(f)$ is not proportional to $K(f)e^{-\kappa U(f)}$ for some κ. So in a certain sense, such a "first-principles" argument for a prior is not self-consistent. In particular, the first principles arguments which have been offered in favor of the so-called "entropic prior" but which do not fix γ (e.g., (Skilling 1989)) suffer from this problem. As another example, with $U(f) = -\log[V(f)]$ theorem one implies that a Dirichlet prior with an unspecified exponent (i.e., a non-delta function $P(\gamma)$) is not a Dirichlet prior. (A similar point is made in [10].)

Second, if the likelihood is nowhere-zero, theorem one says that there is a non-zero lower bound on the error of using evidence to set the posterior. The only question is how low the bound is. To address this make the definition $P(f \mid D) = P(D \mid f)\,[P_E(f) + Er(f)]\;/\;P(D)$, where "$P_E(f)$" means the evidence procedure's approximation to $P(f)$. So if $P(D) \simeq P_E(D)$, the error in the evidence procedure's estimate for the posterior equals $P(D \mid f) \times Er(f)/P(D)$. Therefore we can have arbitrarily large $Er(f)$ for a particular f and not introduce sizable error into the posterior of that f, but only if the likelihood is small for that f. As D varies, the set of those f whose likelihood is not small varies. And as such a set of f varies, the γ (if there is one) such that for those f $P(f \mid \gamma)$ is a good approximation to $P(f)$ varies. When it works, the $\gamma(D)$ returned by the evidence procedure reflects this changing of γ with D.

In general though, one needn't use the evidence procedure to estimate a posterior, but might instead use it for other purposes (see section 4). To circumvent the issue of how the posterior gets used, we will examine the evidence procedure's error as an estimator of an expectation value $\int df' A(f') \times P(f' \mid D)$, where f' is a dummy f variable, and $A(.)$ is determined by the use we have in mind for the posterior.

For example, $A(f') = f'$ if we're interested in the posterior average f. If we're interested in the posterior directly, then $A(f') = A(f, f') = \delta(f - f')$, and expected A is a function of f as well as f'. As a final example, if we're interested in the predictive distribution, then $A(f') = P(\text{new data set} = D' \mid f')$, and A is a function of D' as well as f'.

To analyze such expectation values, let expressions of the form "$E_f(A$ ... stuff)" mean $\int df' A(f') \times P(f'$ ... stuff), where "stuff" can involve f', conditional bars, or whatever; E_f expectation values are over f alone. So for example $E_f(A \mid D) \equiv \int df' A(f') \times P(f' \mid D)$, and $E_f(A, \gamma \mid D) \equiv \int df' A(f') \times P(f', \gamma \mid D)$. (This is slightly non-standard use of the "$E(.)$" notation.) Also, take expressions like "$P(\gamma^* + \delta, ...)$" to be shorthand for "$P(\gamma = \gamma^* + \delta, ...)$".

The intuition for when the evidence procedure works for expectation values is analogous to the intuition for posteriors; the posteriors intuition is based on equation (2), and the expectation values intuition is based on the very similar equation

$$E_f(A \mid D) \;=\; \int d\gamma \frac{E_f(A, \gamma \mid D)}{P(\gamma \mid D)} P(\gamma \mid D) \;\propto\; \int d\gamma E_f(A \mid \gamma, D) P(\gamma \mid D). \qquad (10)$$

Just like equation (3), equation (10) suggests (!) that if $P(\gamma \mid D)$ is sharply peaked about γ^* and $E_f(A \mid \gamma, D)$ is slowly varying, then $E_f(A \mid D) \;\simeq\; E_f(A \mid \gamma^*, D)$.

We now present several theorems which formalize this intuitive reasoning. These theorems give upper and lower bounds on the error induced by using the evidence procedure. In these theorems we never need to specify $A(.)$. In addition, we don't need to assume anything special about the probability distributions, e.g., that they're linear gaussian models.

We will consider three properties:

1) How sharp the γ-peak of $P(\gamma \mid D)$ is.
2) How much $E_f(A \mid \gamma, D) \;=\; E_f(A, \gamma \mid D) / P(\gamma \mid D)$ varies around that peak of $P(\gamma \mid D)$. (This provides the scale for measuring the peakedness of $P(\gamma \mid D)$.)
3) How $E_f(A, \gamma \mid D)$ behaves for γ significantly far from that peak of $P(\gamma \mid D)$. (This - not peakedness of $P(\gamma \mid D)$ - determines if we are justified in ignoring the tails in our integrals.)

Formally, first choose a γ^* and a $\delta \;>\; 0$.

In practice these will usually serve as the peak position and peak width of $P(\gamma \mid D)$ respectively, and we will loosely refer to them as such. (Note though that we make no such stipulations in their definitions, and the theorems presented below don't rely on their serving those functions.)

Our first two definitions characterize the "peakedness" of $P(\gamma \mid D)$; the smaller λ and/or ρ, the more "peaked" the distribution.

$\lambda \equiv$ max $[\frac{P(\gamma^* + \delta) | D)}{P(\gamma^* | D)}, \frac{P(\gamma^* - \delta | D)}{P(\gamma^* | D)}]$;

We will say "condition (i) holds" if λ is small. It is usually assumed that $\lambda < 1$.

$\rho \equiv 1 - \int_{\gamma^* - \delta}^{\gamma^* + \delta} d\gamma P(\gamma \mid D)$;

We will say "condition (i') holds" if ρ is small.

Our next definition characterizes how slowly varying $E_f(A \mid \gamma, D)$ is across the peak; the smaller τ, the more slowly varying $E_f(A \mid \gamma, D)$ is across $[\gamma^* - \delta, \gamma^* + \delta]$.

$\tau \equiv \max |E_f(A \mid \gamma, D) - E_f(A \mid \gamma^*, D)|$ across $\gamma \in [\gamma^* - \delta, \gamma^* + \delta]$;
We will say "condition (ii) holds" if τ is small.

Our next two definitions characterize how much tails over γ matter; the smaller ϵ and/or B, the less those tails matter.

$\epsilon \equiv \; |E_f(A \mid D) - \int_{\gamma^*-\delta}^{\gamma^*+\delta} d\gamma E_f(A, \gamma \mid D)|$;
ϵ is the contribution to $E_f(A \mid D)$ arising from $E_f(A, \gamma \mid D)$ lying outside $[\gamma^* - \delta, \gamma^* + \delta]$.
We will say "condition (iii) holds" if ϵ is small.

$B \equiv \max |E_f(A \mid \gamma, D)|$ across $\gamma \notin [\gamma^* - \delta, \gamma^* + \delta]$;
B measures how big $E_f(A \mid \gamma, D)$ can get when γ is outside of $[\gamma^* - \delta, \gamma^* + \delta]$;
We will say "condition (iv) holds" if B is not too large.

"Evidence's error" is the magnitude of the difference between the full Bayesian answer and the evidence procedure's answer: $|E_f(A \mid D) - E_f(A \mid \gamma^*, D)|$. We will say that "evidence works" if evidence's error is small.

We can now formalize the intuition for when evidence works by writing down an upper bound on evidence's error:

Theorem 2: Evidence's error $\leq \; \epsilon + \tau(1 - \rho) + E_f(A \mid \gamma^*, D) \times |\rho|$.

Proof: $|E_f(A \mid D) - \int_{\gamma^*-\delta}^{\gamma^*+\delta} d\gamma [E_f(A \mid \gamma, D) \times P(\gamma \mid D)]| \; = \; \epsilon$, by definition of ϵ. By the definition of τ, $| \int_{\gamma^*-\delta}^{\gamma^*+\delta} d\gamma [E_f(A \mid \gamma, D) P(\gamma \mid D)] \; - \; E_f(A \mid \gamma^*, D) \int_{\gamma^*-\delta}^{\gamma^*+\delta} d\gamma P(\gamma \mid D)| \;\leq\; \tau \int_{\gamma^*-\delta}^{\gamma^*+\delta} d\gamma P(\gamma \mid D)$. Combining, $|E_f(A \mid D) \; - \; E_f(A \mid \gamma^*, D) \int_{\gamma^*-\delta}^{\gamma^*+\delta} d\gamma P(\gamma \mid D)| \;\leq\; \epsilon + \tau \int_{\gamma^*-\delta}^{\gamma^*+\delta} d\gamma P(\gamma \mid D)$. Therefore $E_f(A \mid \gamma^*, D) - E_f(A \mid D) \leq \epsilon + \tau(1-\rho) + E_f(A \mid \gamma^*, D) \times \rho$. QED.

One can find some sufficiency conditions for evidence to work in the literature. These are specific to certain kinds of distributions, and are derived by evaluating the evidence procedure's answer and the exact answer and seeing if the two differ. Of course, if you can evaluate the exact answer, there's no need for an approximation like the evidence procedure in the first place. In contrast, theorem two provides us with some sets of sufficiency conditions which don't rely on evaluating the exact answer.

For example, if conditions (i'), (ii) and (iii) hold, and $E_f(A \mid \gamma^*, D)$ is not too large, then theorem two tells us that evidence's error is small. (We have no guarantees that it's easy to evaluate whether those conditions hold, of course.) Intuitively, condition (iii) is what lets us restrict attention to the region immediately surrounding the peak of $P(\gamma \mid D)$. Condition

(ii) then tells us that $E_f(A \mid \gamma, D)$ doesn't vary across that region, and can therefore be evaluated at $\gamma = \gamma^*$ and pulled out of the integral. The overall error introduced by the value of that remaining integral is reflected in the $E_f(A \mid \gamma^*, D) \times |\rho|$ term.

Note that this remaining error can be minimized either by having a sharp peak (ρ small) or by having $E_f(A \mid \gamma^*, D)$ - the guess of the evidence procedure - be close to zero. So we don't need to have condition (i') hold (i.e., have $P(\gamma \mid D)$ peaked) for evidence to work. (There are a number of other situations in which the evidence procedure can be justified even though $P(\gamma \mid D)$ is not peaked; see section 6 below.) On the other hand, in section one we saw that peaked $P(\gamma \mid D)$ does not guarantee the accuracy of the evidence procedure. Summarizing, the evidence procedure sometimes works even when $P(\gamma \mid D)$ isn't peaked, and there are also circumstances for which it doesn't work despite $P(\gamma \mid D)$'s being peaked.

All of this notwithstanding, when evidence works in practice usually condition (iii) is met by having ρ small, with $E_f(A \mid \gamma, D)$ staying reasonably bounded for γ outside of $[\gamma^* - \delta, \gamma^* + \delta]$. Formally, $\epsilon \leq B \times \rho$, so that conditions (i') and (iv) give condition (iii). In such scenarios, peakedness of $P(\gamma \mid D)$ does go hand in hand with evidence working.

We now turn to the issue of lower bounds on the error of the evidence procedure. Intuitively, one might think that since γ^* is the "dominant contributing γ", the evidence procedure should work for peaked $P(\gamma \mid D)$ in general. The problem is that one can just as easily argue that the "dominant contributing γ" *for what we are interested in* (namely $E_f(A \mid D)$) is given by $\text{argmax}_\gamma E_f(A, \gamma \mid D)$, not $\text{argmax}_\gamma P(\gamma \mid D)$. After all, $E_f(A \mid D)$ is the γ-integral of $E_f(A, \gamma \mid D)$, not of $P(\gamma \mid D)$. This suggests that for evidence to work, γ^* must (nearly) maximize $E_f(A, \gamma \mid D)$.

Indeed, recall that the intuitive justification of the evidence procedure outlined in equation (10) required that the peaks of $E_f(A, \gamma \mid D)$ and $P(\gamma \mid D)$ nearly coincide, lest τ be too large. This reasoning is formalized in the following theorem, which provides a lower bound on τ based on the peak separation, and which uses the λ measure of peakedness.

Theorem 3: If $E_f(A, \gamma, D)$ does not have a γ-peak somewhere within δ of γ^*, then $\tau \geq E_f(A \mid \gamma^*, D)(1 - \lambda) \,/\, \lambda$.

Proof: By hypothesis $E_f(A, \gamma^*, D)$ has no local maximum in $(\gamma^* - \delta, \gamma^* + \delta)$. Therefore we can't have both $E_f(A, \gamma^* - \delta, D)$ and $E_f(A, \gamma^* + \delta, D)$ less than $E_f(A, \gamma^*, D)$. Without loss of generality, assume $E_f(A, \gamma^*, D) \leq E_f(A, \gamma^* + \delta, D)$. Now examine the ratio expectation values $E_f(A \mid \gamma^* + \delta, D) / E_f(A \mid \gamma^*, D)$, which we can write as the product of ratios $[P(\gamma^* \mid D) / P(\gamma^* + \delta \mid D)] \times [E_f(A, \gamma^* + \delta, D) / E_f(A, \gamma^*, D)]$. By our assumption, the second term in square brackets ≥ 1. However by definition of λ, the first term in square brackets $\geq 1/\lambda$. Therefore $E_f(A \mid \gamma^* + \delta, D) \geq E_f(A \mid \gamma^*, D)/\lambda$, and the difference $E_f(A \mid \gamma^* + \delta, D) - E_f(A \mid \gamma^*, D) \geq E_f(A \mid \gamma^*, D) \times (\lambda^{-1} - 1)$. Using the definition of τ, this means that $E_f(A \mid \gamma^*, D) \times (\lambda^{-1} - 1) \leq \tau$. QED.

In terms of equation (1), large τ means that around $\gamma = \gamma^*$, $E_f(A \mid \gamma, D)$ is *not* slowly varying on the scale of the width of the peak of $P(\gamma \mid D)$. Recall though that if τ is large, then the intuition behind the evidence procedure—that $P(\gamma \mid D)$ "picks out" $E_f(A \mid \gamma, D)$ evaluated at $\gamma = \gamma^*$—is faulty. Formally, if τ is large theorem (2) gives a weak upper bound. And by theorem (3) τ is always large if we have a wide separation between our peaks.

In fact, we can use distance between the peaks to give a *lower* bound on evidence's error, to go with the upper bound of theorem two. To do this, define Γ as the magnitude of the distance between γ^* and that γ-maximum of $E_f(A, \gamma, D)$ which lies closest to γ^*.

Theorem 4: If $E_f(A, \gamma \mid D)$ is non-negative for all γ, it follows that evidence's error $\geq E_f(A \mid \gamma^*, D) \times [\Gamma P(\gamma^* \mid D) - 1]$. Equivalently, it follows that evidence's error $\geq E_f(A \mid D) \times [\, 1 - (1 \,/\, \Gamma P(\gamma^* \mid D))\,]$.

Proof: Since evidence's error is non-negative, if $\Gamma = 0$, the theorem trivially holds. If $\Gamma > 0$, γ^* isn't a maximum of $E_f(A, \gamma, D)$. Accordingly, $E_f(A, \gamma, D)$ must either grow as γ increases past γ^* or as it decreases below γ^*. ("Grow" here is taken to mean "stays level or rises".) Without loss of generality assume it grows as γ increases past γ^*. Then the soonest it could stop growing is at $\gamma = \gamma^* + \Gamma$. Therefore $\int_{\gamma^*}^{\gamma^*+\Gamma} d\gamma E_f(A, \gamma, D) \geq \Gamma E_f(A, \gamma^*, D)$, which implies that $\int_{\gamma^*}^{\gamma^*+\Gamma} d\gamma E_f(A, \gamma \mid D) \geq \Gamma E_f(A, \gamma^* \mid D)$. Recall our hypothesis that $E_f(A, \gamma \mid D)$ is non-negative, which implies that $E_f(A \mid D) = \int d\gamma E_f(A, \gamma \mid D) \geq \int_{\gamma^*}^{\gamma^*+\Gamma} d\gamma E_f(A, \gamma \mid D)$; $E_f(A \mid D) \geq \Gamma E_f(A, \gamma^* \mid D)$. So $E_f(A \mid D) - E_f(A \mid \gamma^*, D) \geq E_f(A \mid \gamma^*, D) \times [\Gamma P(\gamma^* \mid D) - 1]$, which proves the first bound. Now define Δ as the evidence's error and use the fact that $E_f(A \mid \gamma^*, D) \geq E_f(A \mid D) - \Delta$ to convert our lower bound on $E_f(A \mid D)$ to $E_f(A \mid D) \geq \Gamma P(\gamma^* \mid D) \times [E_f(A \mid D) - \Delta]$. Rearranging gives the second bound. QED.

Theorem four provides another reason for why having the γ-peaks far apart is bad for the evidence procedure. (An example of testing the evidence procedure by evaluating the distance between the peaks was presented in figure 2.) Note that theorem four does not mean that a small separation between the peaks implies that evidence works. In fact, it is not even true that evidence working means that the peak separation must be small; the overall multiplicative factor in theorem four might be tiny.

Note that our two peaks are the maximizers over γ of two very similar integrals: $\int df' A(f') P(f', \gamma, D)$ and $\int df' P(f', \gamma, D)$. Accordingly, often if one can evaluate the peak of the evidence, one can also evaluate the peak of $E_f(A, \gamma, D)$, and therefore one can evaluate Γ. So if one can use the evidence procedure, usually one can test its validity. In some cases in fact, it's easier to evaluate the peak of $E_f(A, \gamma, D)$ than it is to evaluate the evidence peak (e.g., for the entropic prior - see [16]). In such circumstances, if one has reason to believe that the evidence procedure is valid (so that Γ must be small), it is easier to evaluate α_{ev} by finding the mode of $E_f(A, \gamma, D)$ than by finding the mode of $P(\gamma \mid D)$.

The need for the peaks to coincide can set strong restrictions on the restrictions on the use of the evidence procedure. For example, take $A(f') = \delta(f - f')$, so that expectation values of A are probabilities of f. Assume $P(\gamma \mid D)$ is quite peaked. Say we want to use the evidence procedure to estimate $E_f(A \mid D) = P(f \mid D)$ for some particular f, $\hat{f}$. Then theorem four tells us that for evidence to work, if $P(\hat{f} \mid D)$ is non-negligible (or equivalently the evidence procedure's prediction $P(\hat{f} \mid \gamma^*, D)$ is non-negligible), then Γ must be quite small for $\hat{f}$, i.e., the peak of $P(\hat{f}, \gamma, D) = P(D \mid \hat{f}, \gamma)P(\hat{f} \mid \gamma)P(\gamma)$ must lie close to γ^* (as measured on the scale of $1/P(\gamma^* \mid D)$). Setting the peaks exactly equal gives us an equation for $\hat{f}$ in terms of D (γ^* being a function of D). In general this equation will have a highly restricted solution for $\hat{f}$, $F(D)$ (i.e., $F(D)$ is a low-dimensional manifold in f-space). For example, in the case of the entropic prior, $F(D)$ is a set of f all sharing the same entropy (that entropy value being set by D). In our gaussians case, $F(D)$ is a set of points all sharing the same $|f|^2$ (where again the precise value is set by D - see theorem four of [20]).

So for sufficiently peaked evidence, unless those f with non-negligible posterior all lie in a highly restricted region ($F(D)$), the evidence procedure is guaranteed to have sizable error for some f. Therefore for sufficiently peaked evidence, if the evidence procedure is to correctly estimate the full posterior, that posterior must be highly peaked (i.e., its support must be confined to a highly restricted region). This in turn usually implies that we're in a likelihood dominated regime - in which case there's little reason to apply Bayesian analysis.

These effects can be envisioned with the help of figure 3. Recall that as N rises, the only effect is that all (!) distributions (over both α and f) become more peaked; the shapes of the distributions and in particular the positions of their peaks do not change. This means that the curves in figure 3 get more peaked—but otherwise do not change—as the evidence gets more peaked (cf. parts b and d of figure 3). Accordingly, as the evidence gets more peaked, the set of f which both have non-zero posterior and which have their posterior well approximated by the evidence procedure becomes tightly restricted. Indeed, that set is empty in part d of figure 3. In fact, of the three β's in figure 3, it is only for the β of part c that the "tightly restricted set of f" doesn't quickly vanish with rising N. Yet it is precisely that value of β in part c that is the largest of those depicted in the figure. This illustrates the fact that when the evidence procedure correctly estimates the full posterior we have high β, and that this effect becomes more pronounced as the evidence becomes more peaked (i.e., as N rises). Rephrasing, things must be likelihood-dominated for the evidence procedure to work, especially when the evidence is peaked.

6 Variations on the sufficiency conditions

There are a number of issues related to upper bounds on evidence's error which were peripheral to the discussion in section five but which deserve mention nonetheless. This section summarizes some of them.

Although it is usually associated with peaked $P(\gamma \mid D)$, there are some scenarios in which the evidence procedure can be used even if $P(\gamma \mid D)$ is not peaked. One such scenario was mentioned in the discussion of theorem two: if $E_f(A \mid \gamma^*, D)$ is small, then we don't need small ρ to get a low upper bound on the evidence procedure's error. Another way to avoid the need for peaked evidence arises if we know the value of $(1 - \rho)$. The idea is to exploit an inequality (established in the proof of theorem two) concerning the "multiplicatively corrected" error: $|E_f(A \mid D) - (1 - \rho)E_f(A \mid \gamma^*, D)| \leq \epsilon + \tau(1 - \rho)$. This relation means that if we multiply the evidence procedure's guess by $1 - \rho$ before using it, then we will incur small error regardless of the value of ρ (so long as τ and ϵ are small). On the other hand, although it shrinks the upper bound of the evidence's error, such post-multiplication of the evidence procedure's guess also raises the lower bound on the error: rather than the bound of theorem four, the bound becomes $E_f(A \mid \gamma^*, D) \times [\Gamma P(\gamma^* \mid D) + \rho - 1]$.

Another way to bypass the need for peaked $P(\gamma \mid D)$ arises if we're interested in the ratio of two expectation values rather than the expectation values themselves. Assume that for some ϕ $|E_f(A \mid D) - \phi E_f(A \mid \gamma^*, D)|$ is bounded by the same small constant Δ for both $E_f(A_1 \mid D)$ and $E_f(A_2 \mid D)$. (E.g., have τ and ϵ small for both A's, take $\phi = 1 - \rho$, and apply the inequality mentioned in the preceding paragraph.) Then the ratio $E_f(A_1 \mid D) \,/\, E_f(A_2 \mid D) = [E_f(A_1 \mid \gamma^*, D) + d_1] \,/\, [E_f(A_2 \mid \gamma^*, D) + d_2]$, where both $|d_1|$ and $|d_2|$ are bounded by Δ/ϕ. As an example choose $A_1 = \delta(f - f_1)$ and $A_2 = \delta(f - f_2)$; if $E_f(A_2 \mid \gamma^*, D) \gg \Delta/\phi$ (note that $\Delta/\phi \leq \tau + \epsilon/(1 - \rho)$), then we can write $E_f(A_1 \mid D) \,/\, E_f(A_2 \mid D] \simeq E_f(A_1 \mid \gamma^*, D) \,/\, E_f(A_2 \mid \gamma^*, D)$, and the evidence procedure accurately approximates the ratio of the two posteriors.

There are other scenarios besides those involving ratios where one isn't directly concerned with "evidence's error" as defined in the preceding sections. Most such scenarios have $A(.)$ a function of f as well as f', so our expectation values are functions of f. (Recall that this is the case when posterior expected $A(.)$ is equivalent to the posterior probability of f, for example.). To avoid confusion, in addressing these scenarios we will write expressions like $E_{f'}(A_f \mid D) \equiv \int df' A_f(f')P(f' \mid D)$; since $A(.)$ is a function of two arguments, the subscript on the "E" is modified to indicate exactly which argument is being marginalized, and a subscript is introduced onto the $A(.)$ to indicate the remaining free variable.

For this kind of $A(.)$ one might wish to measure the accuracy of the evidence procedure over all f, rather than just at one particular f. One way to do this is to evaluate a functional of the two functions $E_{f'}(A_f \mid D)$ and $E_{f'}(A_f \mid \gamma^*, D)$. So for example we might be interested in the least upper bound (over all f) of $|E_{f'}(A_f \mid D) - E_{f'}(A_f \mid \gamma^*, D)|$. Since theorem two holds for any individual f, this least upper bound is bounded above by the quantity $\max_f(\epsilon(f) + \tau(f)(1 - \rho) + E_{f'}(A_f \mid \gamma^*, D) \, |\rho| \,)$ (ϵ and τ have dependence on f through their dependence on $A(.)$). This gives the largest possible gap (across f) between the evidence approximation to the posterior and the correct posterior.

Arguments similar to this least upper bound (lub) one can be used to directly bound $\int df |E_{f'}(A_f \mid D) - E_{f'}(A_f \mid \gamma^*, D)|$. More generally, we can use a bound (however arrived at) on $\text{lub}_f(\, |E_{f'}(A_f \mid D) - E_{f'}(A_f \mid \gamma^*, D)| \,)$ to get bounds on the L^n difference between $E_{f'}(A_f \mid D)$ and $E_{f'}(A_f \mid \gamma^*, D)$ for any n. We illustrate this for the case where $A(f, f') = \delta(f - f')$, so that the expectation value we're examining is the posterior distribution of f.

Define "$L^n(x(f) - y(f))$" to mean the L^n difference between $x(f)$ and $y(f)$. We can bound this difference as follows.

Theorem 5: Let μ be an upper bound on $\text{lub}_f(\, |P(f \mid D) - P(f \mid \gamma^*, D)| \,)$. Then $L^n[P(f \mid D) - P(f \mid \gamma^*, D)] \leq \mu \times [2/\mu]^{1/n}$.

Proof: Let W be the volume of the region in f space where $P(f \mid D) \geq \mu$. Let $C \leq 1$ be the integral of $P(f \mid D)$ over the region corresponding to W. Write $P(f \mid D)$ as $h(f)$ and $P(f \mid \gamma^*, D)$ as $g(f)$, for simplicity. Note that both h and g are positive definite and normalized to 1.

Define $V \equiv [1 - (C - W\mu)] \, / \, \mu$. Since $C - W\mu \leq 1$, $V > 0$. Now choose a region of volume V across which $h(f)$ is infinitesimal. We can always do this since $h(f)$ is normalized to 1 and positive definite, and because f space is infinite; there are regions of arbitrarily large volume over which $h(f)$ is arbitrarily small.

Fix h. Make the hypothesis that the g which maximizes $L^n[h(f) - g(f)]$ equals $[h(f) - \mu]$ across the region corresponding to W, has the value μ across the region corresponding to V, and is zero everywhere else.

This $g(f)$ is positive definite and normalized to 1. Furthermore, for this $g(f)$, the value of $L^n[h(f) - g(f)]$ is bounded by $[\mu^n W + \mu^n(1 - (C - \mu W))/\mu + \mu^n(1 - C)/\mu]^{1/n}$, where the last term inside the square brackets is an upper bound on the contribution of the region where $g(f)$ equals zero but $h(f)$ need not. Rewriting this we get $L^n[h(f) - g(f)] \leq \mu[2W + (2 - 2C)/\mu]^{1/n} \leq \mu[2/\mu]^{1/n}$, as in the statement of the result.

Therefore we only need to prove that our hypothesized g is the one which maximizes the L^n difference between $h(f)$ and $g(f)$. To do this, note that we have three regions of interest; call them (i), (ii), and (iii). Region (i) is the region with volume W. Region (ii) is the region having "volume $[\phi - (C - W\mu)]/\mu$ across which $h(f)$ is infinitesimal". Region (iii) is the rest of f-space, across which $g(f)$ has value 0, although $h(f)$ need not.

If our hypothesized g were not the L^n-maximizing g, then it would be possible to increase $L^n[g(f), h(f)]$ by appropriately shifting some of g between the three regions. Now g can't shrink in region (iii) (it's positive definite), and it can't shrink in region (i) (it must lie

within μ of h by definition of μ). Therefore our only possibility is that g increase-or-stays-the-same in both regions (i) and (iii), and decreases-or-stays-the-same in region (ii).
If g increases in region (i) but stays the same in region (iii), to preserve normalization it must decrease in region (ii). For $n \geq 1$, to have $g(f)$ extremize the L^n norm while not exceeding h by more than μ, we want $g(f)$ to equal 0 or μ everywhere in region (ii). Using this, normalization of g, and the fact that $g(f)$ can't differ from h by more than μ across region (i), we see that the L^n norm can only decrease in this "shifting" procedure.
The same result holds if g increases in region (iii) but not in region (i), or if g increases in both those regions. Therefore g can not change without shrinking the L^n norm. QED.

7 Acknowledgements

DHW worked under the auspices of the Santa Fe Institute and was supported in part by NLM grant F37 LM00011. CEMS (cems@lanl.gov) worked under the auspices of Los Alamos National Laboratory. We thank David Wolf, Tim Wallstrom and Richard Silver for helpful discussion. Send correspondence to DHW at The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM, 87501, USA (dhw@santafe.edu).

References

[1] J. Berger, "Statistical Decision Theory and Bayesian Analysis," Springer-Verlag, 1985.

[2] L. Breiman, "Stacked regression," *University of California, Berkeley, Dept. of Statistics*, TR-92-367, 1992.

[3] W. Buntine, A. Weigend, "Bayesian back-propagation," *Complex Systems*, Vol. 5, p. 603, 1991.

[4] A.R. Davies, R.S. Anderssen, R.S., "Optimization in the regularization of ill-posed problems," *J. Australian Math. Soc. Ser. B*, Vol. 28, p. 114, 1986.

[5] D. L. Donoho et al., "Maximum entropy and the nearly black object," *J. R. Stat. Soc. B* , Vol. 54, pp: 41-81, 1992.

[6] R. Duda, P. Hart, "Pattern classification and Scene analysis," Wiley and Sons, 1973.

[7] G. Demoment, "Image reconstruction and restoration: overview of common estimation structures and problems," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 37, pp: 2024-2036, 1989.

[8] N. Fortier et al., “GCV and ML methods of determining parameters in image restoration by regularization: fast computation in the spatial domain and experimental comparison,” *Journal of visual communication and image representation*, Vol. 4, pp: 157-170, 1993.

[9] S. Gull, “Developments in maximum entropy data analysis,” in “Maximum-entropy and Bayesian methods,” J. Skilling (Ed.). Kluwer Academics Publishers, 1989.

[10] E. Jaynes, “Monkeys, kangaroos and alpha,” in “Maximum-entropy and Bayesian methods,” Kluwer Academic Publishers, 1988.

[11] D.J.C. MacKay, “Bayesian Interpolation,” “A Practical Framework for Backpropagation Networks,” *Neural Computation*, Vol. 4, pp: 415 and 448, 1992.

[12] D.J.C. MacKay, “Bayesian non-linear modeling for the energy prediction competition,” In preparation, 1993.

[13] B.D. Ripley, “Statistical Aspects of Neural Networks,” In “Networks and Chaos – Statistical and Probabilistic Aspects,” O.E. Barndorff-Nielsen et al. (Eds.) Chapman and Hall, 1993.

[14] S. Sibisi, “Regularization and inverse problems,” In “Maximum-entropy and Bayesian methods,” J. Skilling (Ed.). Kluwer Academics publishers. 1989.

[15] J. Skilling, “Classic maximum entropy,” In “Maximum-entropy and Bayesian methods,” J. Skilling (Ed.). Kluwer Academic publishers. 1989.

[16] C.E.M. Strauss, D.H. Wolpert, and D.R. Wolf, “Alpha, Evidence, and the Entropic Prior,” in “Maximum-entropy and Bayesian methods,” A. Mohammed-Djafari (Ed.). Kluwer Academics publishers, 1993.

[17] A. M. Thompson et al., “A study of methods of choosing the smoothing parameter in image restoration by regularization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13, pp: 326-339, 1991.

[18] A.M. Thompson, J. Kay, “On some Bayesian choices of regularization parameter in image restoration,” Technical Report from The University of Edinburgh, no number.

[19] G. Wahba, “A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem,” *The Annals of Statistics*, Vol. 13, pp: 1378-1402, 1985.

[20] D.H. Wolpert, “On the use of evidence in neural networks,” In “Advances in Neural Information Processing Systems 5”, Giles et al. (Eds.), Morgan Kauffman Publishers, 1993.

[21] D.H. Wolpert, “Bayesian backpropagation over I-O functions rather than weights,” In “Advances in Neural Information Processing Systems 6”, Cowan et al. (Eds.), Morgan Kauffman Publishers, 1994.